Supplementary Information for: The Application and Evaluation of a Two-Concept Diagnostic Instrument with Students Entering College General Chemistry

Keily Heredia[†], Xiaoying Xu[†], Jennifer E. Lewis*
Department of Chemistry, University of South Florida
4202 E. Fowler Avenue CHE205A, Tampa, Florida 33620
*Corresponding author E-mail: jennifer@usf.edu

Course Description: Preparatory Chemistry

During Fall 2009, the Preparatory Chemistry syllabus comprised thirteen textbook chapters (Tro, 2009) covering the following chemistry topics: matter, energy, atoms, elements, molecules, compounds, chemical composition, chemical reactions, electrons in atoms, chemical bonding, gases, liquids, solids, intermolecular forces, and solutions. Students enrolled in the course during Fall 2009 met in a large lecture section three times a week for 50 minutes. Additionally, students met weekly in smaller sections, supervised by graduate teaching assistants, to complete an in-class activity known as an "investigation." These activities were written by the course instructor with the help of a graduate teaching assistant.

Three groups within the study sample

A total of 364 students out of 725 were found to have taken Preparatory Chemistry (and are identified as Group A). 361 out of 725 students had not taken the Preparatory Chemistry course; however, 125 (34.6%) of them had registered in a previous General Chemistry I course, while 236 (65.4%) had not. Therefore, 34.6% of the students are re-taking General Chemistry I (and are defined as Group B, repeaters). The other 65.4% of the students are taking General Chemistry I for the first time (and are identified as Group C, first-timers). Since repeaters form a large portion of the sample, it is better to keep them as a separate group, rather than combine groups B and C.

Almost all of the 125 repeaters took General Chemistry I very recently. A total of 107 repeaters took the course during Fall 2009, while 18 repeaters took it during Spring 2009. The most recent obtained grade for the majority of these students from General

Chemistry I is C- or below, or a "W" (41 repeaters withdrew from the course during the semester). Only 5 of the repeaters had originally achieved a course grade as high as C or C+ despite deciding to retake the course. These results are as expected, because students need a grade of C or above to proceed to more advanced chemistry courses. For the 364 students with Preparatory Chemistry (Group A), 333 students were enrolled in the Preparatory Chemistry course during the Fall 2009 semester. Twenty-two were enrolled either during Spring 2009 (N=11) or Fall 2008 (N=11). The other nine students took Preparatory Chemistry more than two years ago. The latest obtained score from previous Preparatory Chemistry is C or better for 350 students, and is C- or below including withdrawals for only fourteen students. Due to the small number of repeaters (8.5%), this group was not divided and was kept within Group A, and identified as with Preparatory Chemistry.

For comparison purposes, demographic information is organized by group. Table S1 presents the sex and race/ethnicity for the three groups. In order to have a better understanding about these three groups of students, a chi square (χ^2) statistic is used. A chi-square statistical analysis was first performed in SAS 9.1 to investigate whether the distribution of students by sex is significantly different for these groups. Results showed that the difference in the distribution is significant, χ^2 (2, N = 725) = 20.7, $p < .01$. A second chi-square was performed to examine whether the differential distribution of the students by race/ethnicity is significant. Results for this second analysis were also significant, χ^2 (12, N = 725) = 52.2, $p < .01$. These results mean that students within these three groups are not distributed similarly with respect to sex and race/ethnicity. For example, the group of students with Preparatory Chemistry is composed of more females (68%) than males (32%), and has fewer white students. This group of students also has a more diverse population including a sizable number of underrepresented minority students.

Table S1. Demographics: number of students by sex and race/ethnicity for each group (N = 725)

	With Prep-Chem		Without Prep-Chem			
			Repeaters		First-timers	
	(Group **A**: n = 364)		(Group **B**: n = 125)		(Group **C**: n = 236)	
	No. of students	Percentage	No. of students	Percentage	No. of students	Percentage
Sex						
Female	248	68.1	68	54.4	119	50.4
Male	116	31.9	57	45.6	117	49.6
Race/Ethnicity						
White (Not of Hispanic Origin)	169	46.4	78	62.4	155	65.7
Hispanic	72	19.8	22	17.6	29	12.3
Black (Not of Hispanic Origin)	62	17.0	7	5.6	21	8.9
Asian or Pacific Islander	52	14.3	13	10.4	12	5.1
American Indian or Native Alaskan	3	.8	1	.8	3	1.3
Ethnicity Unspecified	5	1.4	4	3.2	13	5.5
Ethnicity Specified as “Other”	1	.3	0	0	3	1.3

Prior math achievement for this study consists of the score on the quantitative portion of the SAT, which is available for the majority of students in each group. Table S2 presents average SAT Math scores, which are 524, and 559, and 565 for groups A, B, and C respectively. Levene's test found that the variances for each group are not equal, which means an unmodified ANOVA is not a robust way to check for differences among the groups. Therefore, a Welch test was performed (F (2, n = 576) = 22.6, p <.01) to establish difference. The follow-up Tukey test revealed that students without Preparatory Chemistry have higher SAT Math scores than students with Preparatory Chemistry, while

there is no evidence that the average SAT Math score for the repeaters and the first-timers is different.

Table S2 Demographics: SAT Math score for each group (N = 576)

Group	n	Mean	Std	Skewness	Kurtosis	min	max
A. With Prep-Chem	321	524.2	57.4	0.32	0.70	340	710
B. Without: Repeaters	101	559.0	79.6	-0.54	0.20	310	720
C. Without: First-timers	154	565.3	75.3	-0.02	-0.32	380	760

Table S3 shows student responses to a survey taken on the first day of General Chemistry I. Overall, students without Preparatory Chemistry (Groups B and C) are in their first and second year of college, and students with Preparatory Chemistry (Group A) are mostly in their first year of college. In addition, students without Preparatory Chemistry reported having Algebra and Pre-calculus as their highest level of math, while students with Preparatory Chemistry reported having Algebra as their highest level of math. This information shows that the three groups of students enrolled in General Chemistry I are not the same with respect to their previous academic background.

Table S3 Demographics: Academic background[a] (N = 689)

	With Prep-Chem	Without Prep-Chem	
Category	A: (n = 354)	B: Repeaters (n = 111)	C: First-timers (n = 224)
Year in College	1st year (n = 276, 78.0%)	1st year (n = 46, 41.2%)	1st year (n = 73, 32.6%)
High School Chemistry	2 semesters (n = 194, 54.8%)	2 semesters (n = 57, 51.4%)	2 semesters (n = 108, 48.2%)
Highest Level of Math	Algebra/trigonometry (n= 192, 54.2%)	Algebra/trigonometry (n = 32, 28.8%), pre-calculus (n = 41, 36.9%)	Algebra/trigonometry (n =76, 33.9%), pre-calculus (n = 66, 29.5%)

[a] the most frequent answer for items 1, 4, and 5 is provided. The survey items are

1. How many years (including this one) have you attended a college or university?
   a) 1st year b) 2nd year c) 3rd year d) 4th year e) more than 4 years

4. How much chemistry did you have in high school?
   a) No chemistry in high school b) 1 semester c) 2 semesters
   d) 3-4 semesters e) 5 or more semesters
5. Which best describes the highest level of math you've completed?
   a) I have not taken any math courses as advanced as algebra.
   b) algebra and/or trigonometry c) pre-calculus d) calculus I e) calculus II

Factor Analysis

A confirmatory factor analysis (CFA) is useful to estimate how well the 2-concept model for the instrument fits the data gathered with this sample. As is customary for non-continuous data, a robust weighted least square mean and variance approach (WLSMV) was employed to estimate goodness of fit model based on the tetrachoric correlation matrix for the 10 items. In Mplus 5.2, CFA provides parameter estimates and factor loadings as well as information about the misfit of the items. There are general rules to estimate if the proposed model can be considered a good fit to the data. A non-significant chi-square ($p > .05$) suggests an excellent model fit. However, models produced from a large number of scores are likely to have an inflated chi-square value. Therefore, reporting just the chi-square value can be misleading, and additional fit statistics need to be examined. For example, a Comparative Fit Index (CFI) of .95 or higher is often used. The CFI compares the declared model with a baseline model in which none of the items are related. The obtained CFI value then estimates how much better the proposed model is. Additionally, for categorical data, to examine how close the proposed model is to the data, a Weighted Root Mean Square Residual (WRMR) smaller than 1.0 indicates a good fit (Brown, 2006). In addition, factor correlation and item loading are also important for evaluating model fit. A high factor correlation indicates overlapping scales. This means that the scales may be measuring the same thing, and some of the items are thus redundant. Low item loadings (see main manuscript) also indicate that the variance in the items is not represented well, and therefore, these items may be good candidates for revision or elimination.

Before the analysis was performed, data sample size was checked to see if it was within the suggested item to sample ratio for CFA. Researchers recommend a ratio of five or ten respondents for every item for factor analysis. Since the Diagnostic Instrument contains 10 items, a sample size of 679 safely exceeds the recommended respondent to item ratio.

The estimation of the 2-factor model fit is: χ^2 (N = 679, df = 28, p = .66) = 24, CFI = 1.0, WRMR = .69. However, the factor loadings are quite low in many cases, and the correlation between factors is .83, which suggests redundancy. Accordingly, a 1-factor model was also investigated. For that model, χ^2 (N = 679, df = 29, p = .63) = 25.8, CFI = 1.0, WRMR = .69, and the factor loadings are very similar to the 2-factor model. For this particular data set, there is insufficient evidence to conclude that the instrument's two designed factors, particulate nature of matter and chemical bonding, are functioning as discrete factors in the sample, so it is best not to use factor scores for interpretation.

ANCOVA

Since the average SAT Math score for these three groups of students is significantly different, and SAT Math score is known to have a strong correlation with chemistry achievement in this setting, it is important to consider the students' SAT Math scores to make a fair comparison. To provide a close examination of the difference in performance on the diagnostic instrument for the three groups, an ANCOVA analysis was used to account for prior math achievement.

ANCOVA belongs to the category of multiple regressions and aims to examine the relationship between a continuous outcome variable and a categorical predictor while controlling for continuous predictors. By including the covariate in the regression model, ANCOVA has the advantage of reducing the error term and increasing statistical power to find a difference related to the grouping variable. It is useful when the primary interest is in the categorical predictor, and the research question is whether there is a difference in the means of groups.

Before beginning the analysis, the possible violations of assumptions for ANCOVA were checked. These assumptions include those associated with general linear regression methods. For example, one assumption is a linear correlation between the dependent variable and the covariate, which can be examined in a visual way via a scatter plot. Moreover, ANCOVA assumes there is no interaction effect between the categorical and continuous predictor, so this assumption also must be checked.

To determine whether group membership was associated with a difference in performance on the Diagnostic Instrument, ANCOVA was performed in SAS 9.1, using

the GLM procedure with two independent variables. The continuous variable SAT Math score, to control students' prior math achievement, served as the covariate. The other is the grouping variable, which has three levels, signifying membership in Group A, B, or C. The dependent variable is the Diagnostic Instrument score. The distribution for each variable was checked for normality, and the scatter plots of each pair of variables were also examined. A model with an interaction included was first run. The results indicated no evidence of significance for the interaction effect ($F(2, n = 544) = .74$, $p = .48$), which means the assumption of no interaction effect is tenable. The final model results are presented in the main body of the paper. Figure 1 provides a graphical display of the ANCOVA analysis.

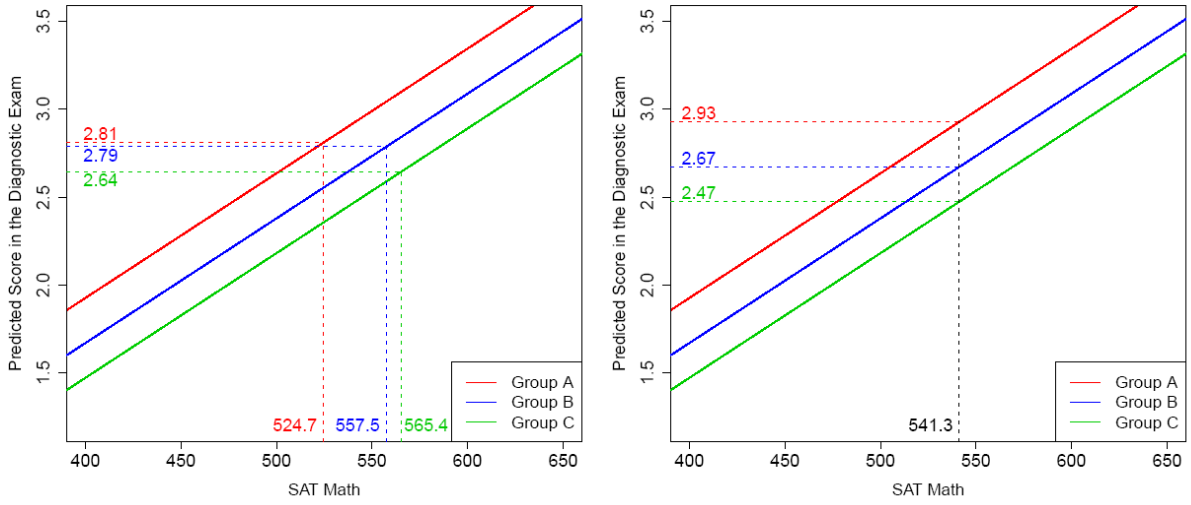

Figure 1. ANCOVA results of the Diagnostic Instrument for three groups when controlling for SAT Math

The unadjusted group mean scores for the diagnostic instrument are 2.81, 2.79 and 2.64 (left side), which are not significantly different from each other. (These mean scores are slighly different from those in Table 2 because the ANCOVA is limited to students with SAT scores.) The red line is for Group A (with prep-chem), the blue line for Group B (repeaters), and the green line is Group C (first-timers). The means of SAT Math for the three groups are 525, 565, and 558 (vertical lines left side), and the grand mean of SAT Math is 541 (vertical line right side). After controlling for SATM, the adjusted predicted means for the diagnostic instrument are 2.93, 2.67 and 2.47 (right side), which are significantly different.

Alternative Conceptions

At the question level, Tables S4 and S5 show that students tend to perform better on the content part of an item (first tier) than on the reasoning part (second tier), and that

getting both parts of an item correct is difficult. These results are consistent with those reported by Othman, *et al*. (2008). Students may know the correct answer for the content question being asked but have not understood why it is so, which indicates a lack of understanding of the concept. Two notable exceptions to this pattern are Items 3 and 5 where the content is less often correct than the reasoning, but it is still the case that getting both content and reasoning correct for these two items is less common than getting one or the other correct.

Table S4: Percentages of students answering the first tier, second tier, and both tiers correctly in the diagnostic test, Items 1-5, with areas of alternative conceptions addressed by each item (N = 679)

Item number	A: With Prep-Chem (n = 348)			B: Repeaters (n = 108)			C: First-timers (n = 223)			Area of alternative conceptions
	tier 1	tier 2	both tiers	tier 1	tier 2	both tiers	tier 1	tier 2	both tiers	
1. What is in the bubbles that are produced in the boiling water?	61.5	49.6	35.3	52.8	34.3	26.9	59.6	35.4	25.1	Phase changes
2. The total mass of the tube and the solid iodine is 27g. The mass after heating will be.	59.2	49.9	44.8	62.0	52.8	47.2	65.0	52.9	45.7	Phase changes; conservation of matter
3. What would the magnified view show after all the water have evaporated?	46.5	50.0	37.6	37.0	40.7	28.7	40.4	49.3	30.0	Phase changes; conservation of matter
4. Crystals of sugar are placed in a beaker of water. If the mixture is left to stand long without stirring, the sugar crystal can no longer be seen.	57.7	49.2	38.8	57.4	43.5	42.6	53.4	39.0	38.6	Dissolving
5. Which of the above properties would be the same for a sample of solid and one single atom of sulphur obtained from the sample.	15.5	40.1	13.8	24.1	33.3	24.1	28.3	29.1	22.9	Particle attributes

Table S5: Percentages of students answering the first tier, second tier, and both tiers correctly in the diagnostic test, Items 6-10, with areas of alternative conceptions addressed by each item (N = 679)

Item number	A: With Prep-Chem (n = 348)			B: Repeaters (n = 108)			C: First-timers (n = 223)			Area of alternative conceptions
	tier 1	tier 2	both tiers	tier 1	tier 2	both tiers	tier 1	tier 2	both tiers	
6. There are molecules in sodium chloride.	30.8	30.1	7.47	21.3	11.1	8.3	24.7	9.9	5.4	Ionic lattice
7. Carbon dioxide has low melting and boiling points.	64.4	28.9	6.61	63.9	10.2	8.3	59.6	7.6	3.6	Bonding: phase changes
8. Calcium fluoride can conduct electricity when molten.	70.4	40.6	19.3	63.0	22.2	20.4	59.2	17.0	16.6	Bonding; ionic lattice; electrical conductivity
9. Solid NaCl does not conduct electricity. However, when sodium chloride is dissolved in water, the resulting solution is able to conduct electricity.	77.3	49.9	37.6	77.8	47.2	38.9	80.7	47.1	36.8	Dissolving; electrical conductivity
10. H_2O and H_2S have similar chemical formulae and structures. At room temp., water is a liquid and hydrogen sulphide is a gas. This difference in state is due to:	72.1	49.6	45.9	66.7	40.7	31.5	78.5	54.3	46.2	Inter-molecular forces; phase changes

References

Brown T. A., (2006) *Confirmatory factor analysis for applied research*, (1st ed.), New York: Guilford Press.

Othman J., Treagust D. F., and Chandrasegaran A. L., (2008) An investigation into the relationship between students' conceptions of the particulate nature of matter and their understanding of chemical bonding, *Int. J. Sci. Educ.*, **30**(11) 1531-1550.

Tro N. J., (2009) *Introductory chemistry*, (3rd ed.), Upper Saddle River, NJ: Pearson.